The Annals of Statistics
1993, Vol. 21, No. 4, 2108–02137

NONPARAMETRIC BINARY REGRESSION: A BAYESIAN APPROACH

By P. Diaconis[1] and D. A. Freedman[2]

Harvard University and University of California

The performance of Bayes estimates are studied, under an assumption of conditional exchangeability. More exactly, for each subject in a data set, let ξ be a vector of binary covariates and let η be a binary response variable, with $P\{\eta = 1|\xi\} = f(\xi)$. Here, f is an unknown function to be estimated from the data; the subjects are independent, and satisfy a natural "balance" condition. Define a prior distribution on f as $\Sigma_k w_k \pi_k / \Sigma_k w_k$, where π_k is uniform on the set of f which only depend on the first k covariates and $w_k > 0$ for infinitely many k. Bayes estimates are consistent at all f if w_k decreases rapidly as k increase. Otherwise, the estimates are inconsistent at $f \equiv 1/2$.

1. Introduction. To illustrate the topic of this paper in a specific context, consider a clinical trial. Each subject has a response variable η and covariates ξ. The response variable is 1 or 0, corresponding to success or failure. For instance, $\eta = 1$ if the subject survives to the end of the study period, else $\eta = 0$. The covariates are a sequence of 0's and 1's. For instance, ξ_1 might be 1 if the subject is male, 0 if female; ξ_2 might be 1 if the subject has high blood pressure, otherwise 0; and so forth. (For present purposes, assignment to treatment or control is just another covariate.)

Given the covariates, assume that the response variables are independent across subjects and

$$P\{\eta = 1|\xi\} = f(\xi). \tag{1.1}$$

Here, f is a measurable function from the space of sequences of 0's and 1's to the closed unit interval [0, 1].

The function f is an infinite-dimensional parameter to be estimated from the data by Bayesian methods. There is a fairly conventional prior distribution which is "nested" or "hierarchical." Begin with a prior π_k supported on the class of functions f that depend only on the first k covariates, so $\xi_{k+1}, \xi_{k+2}, \dots$ do not matter in (1.1). Then treat k as an unknown "hyperparameter," putting prior weight w_k on k. Thus, our prior is of the form

$$\pi = \sum_{k=0}^{\infty} w'_k \pi_k \Big/ \sum_{k=0}^{\infty} w_k, \tag{1.2a}$$

Received February 1991; revised July 1992.

[1]Supported by NSF Grant DMS-86-00235.

[2]Supported by NSF Grant DMS-92-08677 and the Miller Institute.

AMS 1991 *subject classifications*. 62A15, 62E20.

Key words and phrases. Consistency, Bayes estimates, model selection, binary regression.

where

$$(1.2b) \qquad w_k > 0 \quad \text{for infinitely many } k \text{ and } \sum_{k=0}^{\infty} w_k < \infty.$$

The question is whether the Bayes estimates are consistent: Do the posterior distributions pile up around the true f? (More precise definitions will be given shortly.)

Let C_k be the set of strings of 0's and 1's of length k. The prior π_k is defined by the joint distribution it assigns to 2^k parameters, θ_s: $s \in C_k$. Here, θ_s is the probability of success for subjects whose covariate string begins with s. For the present, these θ_s are taken as independent with respect to π_k and uniformly distributed over $[0, 1]$: as we say, "π_k is uniform." Other distributions for θ_s will be considered below. This completes the definition of the prior.

Turning to the data, at stage n there are 2^n subjects indexed by $t \in C_n$. Each subject t has a response variable $\eta_t = \eta(t)$, and an infinite sequence of covariates $\xi_1(t), \xi_2(t), \ldots$. However, the design is "balanced": among the first n covariates, each possible pattern appears exactly once. More specifically,

$$(1.3) \qquad \xi_j(t) = t_j \quad \text{for all } t \in C_n.$$

The remaining covariates $\xi_j(t)$ for $j > n$ are uniform and independent. Call this data structure a "balanced design of order n." The assumptions are made to simplify the calculations below. The designs can be nested in an obvious way, by adding 2^n subjects to go from stage n to stage $n + 1$, but the joint distribution of the designs for various n's will not matter.

Before stating the main theory, we give a more careful definition of consistency. Let $C_\infty = \{0, 1\}^\infty$; so $x \in C_\infty$ has coordinates $x_1, x_2, \ldots$ which are 0 or 1. Write λ^∞ for the uniform measure on C_∞, that is, Lebesgue measure. With respect to λ^∞, the coordinates are independent, and $\lambda^\infty\{x_j = 1\} = 1/2$. By definition, the parameter space Θ is the set of measurable functions from C_∞ to $[0, 1]$; functions which are equal a.e. are identified. Put the L_2 metric on the parameter space. Of course, all the L_p metrics on Θ give rise to the same topology for $1 \le p < \infty$, as does convergence in measure—by the dominated convergence theorem. We write $\|\cdot\|_p$ for the L_p norm.

A typical neighborhood $N(f, \delta, \varepsilon)$ of f will be defined in Definition 1.4. More formally, the $N(f, \delta, \varepsilon)$ are a basis for the neighborhoods at f. [Using weak rather than strong inequalities in Definition 4 is an arbitrary choice.]

(1.4) Definition. If $f \in \Theta$ and $\delta, \varepsilon > 0$, let $N(f, \delta, \varepsilon)$ be the set of $h \in \Theta$ with

$$\lambda^\infty\{x: x \in C_\infty \text{ and } |h(x) - f(x)| \le \varepsilon\} \ge 1 - \delta.$$

If π is a prior probability on Θ, the posterior probability $\tilde{\pi}_n$ on Θ is the conditional law of f given the data; this will be computed explicitly in Section 2. By definition, the prior π is "consistent at f" if $\tilde{\pi}_n\{N(f, \delta, \varepsilon)\} \to 1$ almost surely as $n \to \infty$, provided the data are generated according to a sequence of balanced designs and (1.1) obtains, so f is the true value of the parameter. This frequentist notion of consistency, and its role in Bayesian inference, is discussed in Diaconis and Freedman (1986). The main theorem of this paper can now be stated.

(1.5) THEOREM. *Suppose the designs are balanced, π_k is uniform for all k, the prior π is hierarchical in the sense of* (1.2), *and $f \not\equiv 1/2$. Then π is consistent.*

The case $f \equiv 1/2$ is covered by Theorem 1.9. For example, suppose the π_k are uniform and $w_k = r^k$ for $k \geq 0$. Then π is consistent at all f if $r < \sqrt{1/2}$; but π is inconsistent at $f \equiv 1/2$ if $r > \sqrt{1/2}$.

De Finetti (1959, 1972) studied the performance of Bayes estimates where the data are exchangeable given covariates; also see Bruno (1964). This paper gives precise results in a version of this problem. What is the connection between Theorem 1.5 and the de Finetti's work? From his perspective, subjects with the same covariates would be of the same type and exchangeable. In the present setup, "theory 0" says that all subjects are exchangeable, that is, of the same type. "Theory 1" says that all subjects with $\xi_1 = 0$ are exchangeable, as are all subjects with $\xi_1 = 1$, but the two groups are not exchangeable: so there are two types of subjects. And, so forth. De Finetti studied an example with only three types, and found that the Bayes estimates converged very slowly to the true parameters. (He did use the frequentist notion of consistency as a benchmark.)

In the present set-up, a Bayesian who believes theory k would have prior π_k: subjects would be of the same type provided their first k covariates agreed; in all, there would be 2^k types. A balanced design of order $n > k$ would provide a chance to observe 2^{n-k} subjects for each of the 2^k types. And within a type, the response variables would indeed be exchangeable. However, if $w_0 \gg w_1 \gg \cdots$, it takes a long time for the data to swamp the prior: the posterior tends to concentrate on theories with too few types of subjects. That was the content of de Finetti's example.

It is natural to conjecture that with infinitely many types, and rapidly decreasing w_k, the data may never swamp the prior, so Bayes estimates would be inconsistent. The facts are otherwise. If w_k decreases rapidly, the Bayes estimates are consistent. In the present setup, there are a continuum of types because there are countably many covariates. The prior π says there are only finitely many types, although that number can be indefinitely large. Consistency is all the more surprising.

More curious still, if w_k decreases slowly, and the π_k are uniform, Bayes estimates can be inconsistent—for the function which is identically 1/2. This f is the mean of π; and no covariates matter, so there is only one type of subject in the clinical trial. Of course, the Bayesian statistician does not know this a priori, and the "curse of dimensionality" strikes again.

Coming back to the mathematics, we establish results on consistency and inconsistency for a more general class of priors with "Γ-uniform π_k"; these will be defined in Definition 1.7. First, the success probabilities θ_s are defined more carefully in Definition 1.6.

(1.6) DEFINITION. Fix $k \geq 0$. Let $\Theta_k \subset \Theta$ consist of the functions h which depend only on the first k covariates. If $h \in \Theta_k$, then $\theta_s(h)$ is the value of $h(x)$ when $x \in C_\infty$, $s \in C_k$ and $x_j = s_j$ for $1 \leq j \leq k$.

Informally, if π_k is Γ-uniform, then π_k envisions 2^k types of subjects, each with a distinct success probability θ_s. The θ_s are independent but not identically distributed: each θ_s has its own prior density γ_s. These γ_s are uniformly bounded above by $B < \infty$, and below by $b > 0$. Furthermore, the mean of γ_s is constrained to be in a given finite subset F of the open unit interval. The index s runs through C_k, the set of strings of 0's and 1's of length k.

To state the formal definition more compactly, each $s \in C_k$ is also viewed as a subset of C_∞:

$$s = \{x\colon x \in C_\infty \text{ and } x_j = s_j \text{ for } 1 \leq j \leq k\}.$$

If $f \in \Theta_k$, then $f(x)$ is constant as x ranges over $s \in C_k$, when s is viewed as a subset of C_∞.

(1.7) DEFINITION. Fix $0 < b < B < \infty$, and a finite subset F of $(0, 1)$. Consider the class Γ of all densities γ on $[0, 1]$, with $b \leq \gamma \leq B$ and $\int_0^1 \theta\gamma(\theta)\,d\theta \in F$. Consider π_k which concentrate on Θ_k, make the 2^k success probabilities θ_s: $s \in C_k$ independent, and give each of them a density γ_s in the class Γ. Let g_s be the mean of γ_s, so $g_s = \int_0^1 \theta\gamma_s(\theta)\,d\theta \in F$. Write $g_k(s) = g_s$, and extend g_k to a function on C_∞ by setting $g_k(x) = g_k(s)$ for all $x \in s$. Assume g_k comes from a limiting function g_∞ that takes values in F and is continuous on C_∞. Of course, a "continuous" function on C_∞ that takes only finitely many values must be piecewise constant on C_k, for all large k. To avoid extraneous complications, suppose that $g_k \equiv g_\infty$ for all $k \geq n_1$. This completes the definition of Γ-uniform π_k.

For comparison, the original setup had $b = B = 1$ and $F = \{1/2\}$, so $g_k \equiv g_\infty \equiv 1/2$ for $k \geq 0$. Theorem 1.5 continues to hold for Γ-uniform π_k: there is consistency at f unless $f = g_\infty$ a.e., as Theorem 1.8 shows. The case $f \equiv g_\infty$ is handled by Theorem 1.9.

(1.8) THEOREM. *Suppose the designs are balanced; the* π_k *are* Γ-*uniform in the sense of Definition* 1.7, *the prior* π *is hierarchical in the sense of* (1.2), *and* $f \not\equiv g_\infty$. *Then* π *is consistent.*

(1.9) THEOREM. *Suppose the designs are balanced; the* π_k *are* Γ-*uniform in the sense of Definition* 1.7; *the prior* π *is hierarchical in the sense of* (1.2); *and*

$f \equiv g_\infty$. *Let* l *be the smallest* k *with* $w_k > 0$. *Let* $\beta = (1/2)\log 2$ *and* $\delta > 0$.

(a) *Suppose* $\sum_{k=n}^{\infty} w_k < \exp(-\beta n 2^l - \delta n 2^l)$ *for all large* n. *Then* π *is consistent at* f.

(b) *Suppose* $\sum_{k=n}^{\infty} w_k > \exp(-\beta n 2^l + \delta n 2^l)$ *for infinitely many* n. *Then* π *is inconsistent at* f.

What happens if π is inconsistent? For $m > 0$, let $\pi_{(m)}$ be the prior π with theories 1 through m deleted. Let $\|\cdot\|$ be the variation norm, and suppose for instance that $w_k = 1/k^2$. Fix K large but finite. Asymptotically, theories indexed by $k \leq n + K$ are negligible. Indeed, $\|\tilde{\pi}_n - \pi_{(n+K)}\| \to 0$ almost surely as $n \to \infty$. This is true for any finite K. In the long run, there are infinitely too many types. And the success probabilities are independent, so the f's you have left are very wiggly indeed.

Suppose f depends on only finitely many covariates, say $\xi_i, \dots, \xi_k$. Under the conditions of Theorem 1.8 or 1.9a, the posterior concentrates on such functions: $\tilde{\pi}_n\{C_k\} \to 1$ a.s. as $n \to \infty$. The argument is about the same as for the theorems. Thus, the Bayesian gets the order of the model right too. This is a bit surprising, because many rules for model selection will over-estimate k.

Section 2 gives explicit formulas for the posterior; Section 3, some preliminary estimates. Theorem 1.8 is proved in Section 4, and Theorem 1.5 is a special case. Theorem 1.9 is proved in Section 5.

Our results may seem a bit special; however, we believe the phenomenon to be fairly general. We think it applies to other sequences of nested models, and other kinds of problems (like regression). For example, see Diaconis and Freedman (1988, 1991); in the latter, we show that very similar results hold for unbalanced data, with random covariates.

Here is another kind of generalization. We have assumed that π_k is Γ-uniform in the sense of Definition 1.7, but the arguments go through almost without change for π_k^* which make the joint distribution of the θ_s absolutely continuous, having a density (in R^{2^k}) relative to π_k, bounded above by $B^* < \infty$ and below by $b^* > 0$, where b^* and B^* do not depend on k. For the proof, let

$$\pi^* = \sum_{k=0}^{\infty} w_k \pi_k^* \Big/ \sum_{k=0}^{\infty} w_k.$$

Then $b^*\pi \leq \pi^* \leq B^*\pi$, and $(b^*/B^*)\tilde{\pi}_n \leq \tilde{\pi}_n^* \leq (B^*/b^*)\tilde{\pi}_n$. Indeed, for any events C and D, $(b^*/B^*)\pi(C|D) \leq \pi^*(C|D) \leq (B^*/b^*)\pi(C|D)$.

Our concern is with the consistency of Bayes estimates. Of course, consistent estimates (based on other principles) are generally available. For example, Stone (1982) gives consistent nearest-neighbor estimates for f and shows that under smoothness conditions, these estimates achieve best possible rates of convergence. Cox and O'Sullivan (1990) derived similar results for penalized likelihood estimates of $\log(f/1 - f)$. O'Sullivan, Yandell and Raynor (1986) describe applications. Leonard (1978) discusses connections between penalized likelihood and Bayesian methods.

There have been many other studies of nonparametric regression, using nested increasing sequences of finite-dimensional approximations. Akaike's criterion was adapted to regression by Shibata (1981). Shibata considers increasing families of regression functions, for instance, all polynomials of degree k_n or less with $k_n = o(n)$ as $n \to \infty$. For each n, a model size $\hat{k}_n \le k_n$ is chosen to minimize estimated prediction error. This estimate is the sum of bias and variance terms. Shibata proves that the bias term is asymptotically smallest with his rule, but he does not address consistency issues. Schwartz (1978) proposed a Bayesian version of model selection when the dimensionality is bounded. Our paper can be viewed as an extension of Schwartz's idea to the infinite-dimensional case. For reviews of the literature on model selection, see Breiman and Freedman (1983), Li (1986) or Shibata (1986).

There is related literature on sieves and orthogonal series. With sieves, one considers an increasing family of finite-dimensional models in an infinite dimensional space. A cut-off sequence $k_n \uparrow \infty$ is chosen. With n data points, one estimates the k_nth model by maximum likelihood as in Geman and Hwang (1982) or least squares as in Cox (1988). Also see Grenander (1981). With appropriate smoothness conditions, k_n can be chosen to get consistency. Cox carries out the details for regression problems. Our paper puts a posterior distribution on k, rather than imposing a sharp cut-off.

In the density-estimation context, orthogonal-series estimators consider $\hat{f}(x) = \sum_{i=0}^{k} \hat{\beta}_i f_i(x)$ for a fixed series of orthogonal functions $\{f_i\}$. The weights $\hat{\beta}_i$ are estimated from the data. The order k can be chosen by cross validation, as suggested by Rudermo (1982) and Bowman (1984). For reviews, see Hall (1987) or Eubank (1988). Our Bayes estimates are formally similar, being infinite mixtures of finite-dimensional Bayes estimates, with data-driven weights.

Our consistency proof shows that the prior piles up around the MLE, which is consistent. There are similar ideas in Datta (1991) and Gilliland, Hannan and Huang (1976). Of course, LaPlace (1774) deserves mention too.

2. Computing the posterior. Fix n, and consider a balanced design of order n. The posterior $\tilde{\pi}_n$ for π will be computed in Lemma 2.14. First, we compute the posterior for π_k with $k \le n$, then for $k \ge n$. To get started, fix $k \le n$. For $s \in C_k$, let X_s be the number of successes among subjects whose covariate sequence begins with s. More formally, $\eta(t)$ is the response for subject $t \in C_n$, and

$$X_s = \sum_{t \in C_n} \{\eta(t) : t_i = s_i \text{ for } i = 1, \dots, k\}. \tag{2.1}$$

Assume that π_k is Γ-uniform in the sense of Definition 1.7, so the success probabilities θ_s are independent as s ranges over C_k, and θ_s has the density $\gamma_s \in \Gamma$. The parameter space is Θ. Let Ω be an underlying probability space, on which the response variables $\eta(t)$ and covariates $\xi_i(t)$ are defined. For $f \in \Theta$, let P_f be the probability on Ω which makes the response variables and covariates distributed in a balanced design so that (1.1) holds.

As usual, π_k can be extended to a probability on $\Theta \times \Omega$, by the formula

$$\pi_k(A \times B) = \int_A P_f\{B\}\pi_k\{df\}.$$

In this formula, A is a measurable subset of Θ and B is a measurable subset of Ω. We endow Θ with the σ-field generated by the strong L_2 topology: $f \to P_f\{B\}$ is measurable because

$$f \to P_f\{\xi_1(t) = e_1, \xi_2(t) = e_2, \dots, \xi_n(t) = e_n, \eta(t) = e\}$$

is continuous, the e's being 0 or 1. Write $\text{bin}(m, \theta)$ for the binomial distribution, with m trials and success probability θ.

(2.2) LEMMA. *Suppose $k \le n$ and π_k is Γ-uniform. With respect to the prior π_k, the pairs (θ_s, X_s) are independent as s ranges over C_k. The parameter θ_s has density $\gamma_s \in \Gamma$. Given θ_s, the number of successes X_s is* $\text{bin}(2^{n-k}, \theta_s)$.

The proof of Lemma 2.2 is omitted as routine. In Lemma 2.2 and similar contexts, π_k is viewed as a probability on $\Theta \times \Omega$. For $\gamma \in \Gamma$, $m = 1, 2, \dots$ and $j = 0, 1, \dots, m$, let

$$\text{(2.3a)} \qquad \gamma(m, j, \cdot)\colon \theta \to \frac{\theta^j(1-\theta)^{m-j}\gamma(\theta)}{\phi(m, j, \gamma)},$$

where the normalizing constant is

$$\text{(2.3b)} \qquad \phi(m, j, \gamma) = \int_0^1 \theta^j(1-\theta)^{m-j}\gamma(\theta)\, d\theta.$$

To intrepret ϕ, suppose a Bayesian with prior density γ on θ tosses a θ-coin m times. Then $\phi(m, j, \gamma)$ is the predictive probability of any particular sequence of outcomes with j heads.

Let $\tilde{\pi}_{k,n}$ be the posterior distribution of f, computed relative to π_k, given the data from a design of order n. Lemma 2.4 computes this posterior for $k \le n$, and is almost immediate from Lemma 2.2.

(2.4) LEMMA. *Suppose $k \le n$ and π_k is Γ-uniform. According to the posterior $\tilde{\pi}_{k,n}$, the success probabilities θ_s are independent as s ranges over C_k, and θ_s has density $\gamma_s(2^{n-k}, X_s, \theta)$ with respect to Lebesgue measure on* $[0,1]$.

Turn now to π_k with $k \ge n$. There are 2^k parameters θ_s, indexed by $s \in C_k$; and $2^n \le 2^k$ subjects indexed by $t \in C_n$. Lemma 2.6 describes the extension of π_k to $\Theta \times \Omega$ for designs of order $k \ge n$. The idea is simple. There are 2^k independent coin-tossing experiments, with random success probabilities. And 2^n of the coins actually get tossed—once each; as we say, there are observations on those parameters. The remaining $2^k - 2^n$ coins do not get tossed at all, and there are no observations on their parameters. The notation is complicated, because we have to keep track of which parameters are which.

According to theory k, covariates beyond the kth do not matter. For subject $t \in C_n$, covariates $n+1, \dots, k$ are denoted $\xi_{n+1}(t), \dots, \xi_k(t)$: these are random. Let τ_t be the first k covariates for subject t, that is,

$$\tau_t = t, \xi_{n+1}(t), \dots, \xi_k(t) \in C_k. \tag{2.5}$$

Let $C_k^* = \{\tau_t \colon t \in C_n\}$, so C_k^* is a random subset of C_k, and $|C_k^*| = 2^n$. Let $C_k^{**} = C_k \setminus C_k^*$, so $|C_k^{**}| = 2^k - 2^n$.

The parameters with observations are indexed by $s \in C_k^*$; the others, by $s \in C_k^{**}$. (The number of observations per parameter is either 1 or 0.) In other terms, C_k^* is the set of k-strings of covariates for subjects in the design of order $n \le k$; C_k^{**} is the set of k-strings of covariates for subjects not in the design: the response η_s has not been observed at stage $n < k$ for $s \in C_k^{**}$, so no distribution is given for η_s in Lemmas 2.6 and 2.4. The proof of Lemma 2.6 is routine, and Lemma 2.7 follows. [If $k = n$, C_k^{**} is empty, and the formulations in (2.2)–(2.4) apply as well.]

(2.6) LEMMA. *Suppose $k \ge n$ and π_k is Γ-uniform. Condition on the covariates for the 2^n subjects. With respect to the prior π_k,*

$$(\theta_{\tau_t}, \eta_t) \colon t \in C_n \quad \text{and} \quad \theta_s \colon s \in C_k^{**}$$

are all independent; θ_s has density $\gamma_s \in \Gamma$ for all $s \in C_k$. For $t \in C_n$, given θ_{τ_t}, the response variable η_t is 1 with probability θ_{τ_t} and 0 with probability $1 - \theta_{\tau_t}$.

(2.7) LEMMA. *Suppose $k \ge n$ and π_k is Γ-uniform. According to the posterior $\tilde{\pi}_{k,n}$, the success probabilities θ_s are independent as s ranges over C_k. If $t \in C_n$, then θ_{τ_t} has density $\gamma_{\tau_t}(1, \eta_t, \theta)$ with respect to Lebesgue measure on $[0,1]$. If $s \in C_k^{**}$, then θ_s has density $\gamma_s \in \Gamma$.*

To compute the posterior relative to π, the π_k-predictive probability of the data is needed. To set up the notation, recall the normalizing constant ϕ from (2.3b). Let

$$\rho_{k,n} = \prod_{s \in C_k} \phi(2^{n-k}, X_s, \gamma_s) \quad \text{for } 0 \le k \le n. \tag{2.8}$$

Recall τ_t from (2.5). Let

$$\rho_{k,n} = \prod_{t \in C_n} \phi(1, \eta_t, \gamma_{\tau_t}) \quad \text{for } k \ge n. \tag{2.9}$$

By Lemmas 2.2 and 2.6, $\rho_{k,n}$ is the π_k-predictive probability of the data.

Before going on to compute the posterior relative to π, we pause to rewrite (2.9) in terms of entropy. Recall from Definition 1.7 that the prior means fit together into the function g_∞, which is constant on each $t \in C_n$, provided $n \ge n_1$. Write $g_\infty(t)$ for the common value of $g(x)$ when $x \in C_\infty$ but $x_j = t_j$ for $1 \le j \le n$.

Define the relative entropy function $H(p, \theta)$ as usual:

$$H(p, \theta) = p \log \theta + (1-p) \log(1-\theta), \tag{2.10}$$

unless $p = \theta = 0$ or $p = \theta = 1$. The function H is left undefined at the corners, where it has bad singularities.

(2.11) LEMMA. *Suppose the designs are balanced and the* π_k *are* Γ-*uniform. For all sufficiently large* n, *for all* $k \geq n$,

$$\log \rho_{k,n} = \sum_{t \in C_n} H[\eta_t, g_\infty(t)].$$

PROOF. If η is 0 or 1, and $g_s = \int_0^1 \theta \gamma_s(\theta)\, d\theta$, then $\log \phi(1, \eta, \gamma_s) = \eta \log g_s + (1 - \eta)\log(1 - g_s) = H(\eta, g_s)$. So

$$\log \rho_{k,n} = \sum_{t \in C_n} H(\eta_t, g_{\tau_t}).$$

If $t \in C_n$ and $n \geq n_1$, then g_∞ is constant on t and $g_{\tau_t} = g_\infty(t)$, by Definition 1.7 of Γ-uniformity. □

Turn now to the posterior $\tilde{\pi}_n$, computed relative to π. Informally, the "theory index" k in (1.2) is a parameter, which has a posterior distribution relative to π. Let

$$\tilde{w}_{k,n} = w_k \rho_{k,n}. \tag{2.12}$$

Now, $\pi_k\{\text{data}\}/\pi\{\text{data}\} = \tilde{w}_{k,n}/\sum_{k=0}^{\infty} \tilde{w}_{k,n}$. So

$$\tilde{\pi}_n(k) = \tilde{w}_{k,n}\Big/ \sum_{k=0}^{\infty} \tilde{w}_{k,n}. \tag{2.13}$$

As Lemma 2.14 shows, $\tilde{\pi}_n$ is a mixture of the posteriors $\tilde{\pi}_{k,n}$, with weights equal to the $\tilde{w}_{k,n}$ of (2.12).

(2.14) LEMMA. *Suppose* π *is hierarchical in the sense of* (1.2), *and the* π_k *are* Γ-*uniform. Given the data from a design of order* n, *the posterior is*

$$\tilde{\pi}_n = \sum_{k=0}^{\infty} \tilde{w}_{k,n} \tilde{\pi}_{k,n} \Big/ \sum_{k=0}^{\infty} \tilde{w}_{k,n}.$$

The proof is omitted as routine.

REMARK. The Bayes estimate of f under quadratic loss is just the mixture $\sum_{k=0}^{\infty} \tilde{w}_{k,n} \tilde{f}_{k,n}/\sum_{k=0}^{\infty} \tilde{w}_{k,n}$, where $\tilde{f}_{k,n}$ is the mean of $\tilde{\pi}_{k,n}$. This posterior mean is easily computed. From the point of view of π_k, there are 2^k independent experiments going on, one for each type of subject. These types are indexed by $s \in C_k$. For each type of subject, there are 2^{n-k} tosses of coin, which lands heads with probability θ_s; and π_k puts prior density γ_s on θ_s. So, you compute the posterior mean of γ_s given the number of successes among the subjects of type s. And that is the value of $\tilde{f}_{k,n}(x)$ for x with $x_j = s_j$, $1 \leq j \leq k$.

3. Some estimates. The entropy function H is defined as usual:

$$H(p) = \begin{cases} p\log p + (1-p)\log(1-p), & \text{for } 0 < p < 1, \\ 0, & \text{for } p = 0 \text{ or } 1. \end{cases} \tag{3.1}$$

Recall $\phi(m, j, \gamma)$, the normalizing constant from (2.3b). If $\gamma \equiv 1$, abbreviate $\phi(m, j, \gamma)$ to $\phi(m, j)$. Then

$$\phi(m, j) = \frac{j!(m-j)!}{(m+1)!}.$$

The $\phi(m, j, \gamma)$ can be estimated using ϕ^*, defined as follows. For $m = 1, 2, \ldots$ and $j = 0, \ldots, m$, let $\hat{p} = j/m$ and

$$\phi^*(m, j) = \begin{cases} e^{mH(\hat{p})} \cdot \dfrac{1}{\sqrt{m}} \cdot \sqrt{2\pi}\sqrt{\hat{p}(1-\hat{p})}, & \text{for } 0 < j < m, \\ \dfrac{1}{m}, & \text{for } j = 0 \text{ or } m. \end{cases} \tag{3.2}$$

(3.3) LEMMA. *Let* $m = 1, 2, \ldots$. *Let* $\gamma \in \Gamma$, *so* $0 < b \le \gamma \le B < \infty$.

(a) *There are* $0 < a < A < \infty$ *such that for all* $\gamma \in \Gamma$ *and all* $j = 0, 1, \ldots, m$, $a < \phi(m, j, \gamma)/\phi^*(m, j) < A$.

(b) $1/[2^m(m+1)] < \phi(m, j) \le 1/(m+1)$.

(c) $-m < \log \phi(m, j) \le -\log(m+1)$.

(d) $-m + \log b < \log \phi(m, j, \gamma) < 0$.

(e) $\phi(m, j, \gamma) \le B/(m+1)$.

PROOF. Claim (a). Clearly, $b\phi(m, j) \le \phi(m, j, \gamma) \le B\phi(m, j)$. If $j = O(1)$ or $m - j = O(1)$, the result is clear. Now use Stirling's formula on $\phi(m, j)$ for j and $m - j$ large.

Claim (b). Clearly,

$$\phi(m, j) = 1\Big/\left[(m+1)\binom{m}{j}\right] \quad \text{and} \quad 1 \le \binom{m}{j} < 2^m. \tag{3.4}$$

Claim (c). For the upper bound, use (3.4). For the lower bound, $\binom{m}{j}$ takes its maximum when $j = [m/2]$. Let

$$q(m) = (m+1)\binom{m}{[m/2]}e^{-m} \quad \text{for } m = 1, 2, \ldots.$$

By a direct calculation, $q(m)$ decreases as m increases for $m \ge 2$. For $m = 1$ or 2, by another direct calculation, $q(m) < 1$.

Claim (d). The upper bound is clear, since $\phi(m, j, \gamma)$ represents a probability. The lower bound is immediate from (c), because $\gamma \ge b$ as part of the definition of Γ.

Claim (e). $\phi(m, j, \gamma) \le B\phi(m, j) \le B/(m+1)$ because $\gamma \le B$ as part of the definition of Γ, and $\phi(m, j)$ is maximum at $j = 0$ or m. □

REMARK. If γ is smooth and $\varepsilon \le \hat{p} \le 1 - \varepsilon$ for $\varepsilon > 0$, then

$$\log \phi(m, j, \gamma) - \log \phi^*(m, j) = \gamma(\hat{p}) + O\left(\frac{1}{m}\right).$$

We will not need such estimates for proving Theorem 1.8. The constant $\sqrt{2\pi}\sqrt{\hat{p}(1-\hat{p})}$ in (3.2) and a, A in Lemma 3.3a will be absorbed into error terms. What counts is $\exp[mH(\hat{p})]$. For Theorem 1.9, the $\sqrt{m}$ matters too; $\hat{p}$ near 0 or 1 for theories k near n is a more technical nuisance. The bounds in Lemma 3.3a, d and e are uniform in $\gamma \in \Gamma$; this will be used in the proofs. For related expansions of ϕ, see Johnson (1967, 1970) or Ghosh, Sinha and Joshi (1982).

To state the next result, extend $\phi(m, j, \gamma)$ in (2.3b) from integer $j = 0, 1, \dots, m$ to real x in $[0, m]$.

(3.5) LEMMA. *$x \to \log \phi(m, x, \gamma)$ is strictly convex.*

PROOF. The second derivative with respect to x is

$$\int \left\{\log \frac{\theta}{1-\theta}\right\}^2 q(\theta)\, d\theta - \left\{\int \log \frac{\theta}{1-\theta} q(\theta)\, d\theta\right\}^2,$$

where

$$q(\theta) = \theta^x (1-\theta)^{m-x} \gamma(\theta)/\phi(m, x, \gamma).$$

In particular, q is a density and the second derivative is a variance. □

Of course, there are more general results for exponential families; see Lehmann (1983, page 26 ff. Recall the predictive probabilities $\rho_{k,n}$ from (2.8) and (2.9). We will be estimating these by taking logs, so expected values come into the calculation. To set up the notation, for $m = 1, 2, \dots$ let

$$(3.6) \quad \psi(m, p, \gamma) = E\left\{\frac{1}{m} \log \phi(m, X, \gamma)\right\}, \quad \text{where } X \text{ is bin}(m, p).$$

(3.7) LEMMA. *Let $Y = \sum_{i=1}^m \eta_i$, the η_i's being independent and $0-1$ valued with $P\{\eta_i = 1\} = p_i$. Let $(1/m)\sum_{i=1}^m p_i = p$. Then*

$$E\left\{\frac{1}{m} \log \phi(m, Y, \gamma)\right\} \le \psi(m, p, \gamma).$$

PROOF. This follows from Lemma 3.5, by Theorem 3 in Hoeffding (1956). □

(3.8) LEMMA. *Define the entropy function H by (3.1). For all $p \in [0,1]$ and $\gamma \in \Gamma$:*

(a) $-1 + [(\log b)/m] < \psi(m,p,\gamma) < 0$ *for* $m = 1, 2, \dots$.
(b) *For $m = 2, 3, \dots$ there is an $\varepsilon_m > 0$, which does not depend on p or γ, such that*

$$\psi(m,p,\gamma) \le H(p) - \varepsilon_m.$$

PROOF. Claim (a). Use Lemma 3.3d.

Claim (b). For any particular p and γ, we will show

$$\psi(m,p,\gamma) < H(p). \tag{3.9}$$

Indeed, consider two laws P and Q for $X = (X_1, \dots, X_m)$. According to P, the X_i are iid, each being 1 with probability p and 0 with probability $1-p$. Let Q be the predictive probability for X, for a Bayesian who has a prior density γ on p. Now $P \neq Q$ provided $m > 1$, so

$$E_P\{\log Q(X)\} < E_P\{\log P(X)\}.$$

The left-hand side is $m\psi(m,p,\gamma)$; the right-hand side is $mH(p)$. This proves (3.9). Now put the weak star topology on Pr[0, 1], the space of probabilities on [0, 1]. The class Γ is compact, and $\psi(m,\cdot,\cdot)$ is continuous on $[0,1] \times \Gamma$. This proves (b). □

REMARK. When $m = 1$, $\psi(1,g,\gamma) = H(g)$, where $g = \int \theta\gamma(\theta)\,d\theta$; if $p \neq g$, $\psi(1,p,\gamma) < H(g)$. Of course, $\psi(1,p,\gamma) = H(p)$. Intuitively, tossing a coin with a random parameter is the same as tossing an ordinary coin—provided you only toss it once. This may seem like a trivial observation, but it is the root cause of the inconsistency of Bayes estimates in Theorem 1.9.

(3.10) LEMMA. *Let μ and ν be two probabilities on Θ. The variation distance is $\|\mu - \nu\| = 2\sup_A|\mu(A) - \nu(A)|$. Let c and d be positive real numbers. Then*

$$\left\| \frac{c\mu + d\nu}{c+d} - \mu \right\| \le \frac{d}{c+d}\|\mu - \nu\| \le \frac{2d}{c+d}.$$

The routine proof of Lemma 3.10 is omitted. The following calculations are standard, but are included for ease of reference. Recall the entropy function H from (3.1). Since H is strictly convex,

(3.11) $H(p) + H'(p)(x-p) \le H(x)$ for all $x \in [0,1]$, with equality only at $x = p$.

For $p \in (0,1)$ and $x \neq p$, let

$$H_p\colon x \to \frac{H(x) - H(p) - H'(p)(x-p)}{(x-p)^2}. \tag{3.12}$$

Clearly, H_p can be extended to a continuous, positive function on [0, 1], whose value at p is $(1/2)H''(p)$.

(3.13) DEFINITION. Let $H^*(p)$ be the maximum of H_p on $[0, 1]$.

Reorganizing slightly, we get

(3.14) $H(x) \le H(p) + H'(p)(x - p) + H^*(p)(x - p)^2$ for all $x \in [0, 1]$, with equality only at $x = p$.

(3.15) COROLLARY. *Let X be a random variable taking values in the unit interval. Suppose $E\{X\} = p$ and* $\operatorname{var}\{X\} = \sigma^2 > 0$. *Then*

$$H(p) < E\{H(X)\} < H(p) + \sigma^2 H^*(p).$$

4. Proof of Theorem 1.8. Before proving Theorem 1.8, we outline the argument; and a brief review of the notation may be helpful. The parameter space Θ consists of all measurable functions from $C_\infty = \{0, 1\}^\infty$ to $[0, 1]$; functions which are equal a.e. are identified. We put the L_2 metric on Θ, making it complete and separable but not compact. For $f \in \Theta$, f_k will be the conditional expectation of f given the first k covariates: See (4.1).

Let $\Pr(\Theta)$ be the space of probabilities on Θ. Endow $\Pr(\Theta)$ with the weak star topology; for a discussion of weak star topologies, see Parthasarathy (1967). Then π is consistent at $f \in \Theta$ if $\tilde{\pi}_n$ converges a.s. $[p_f]$ to point mass at f. The prior π is defined by (1.2), making the "theory index" k a parameter: k says how many covariates come into the formula (1.1).

We now outline the proof of Theorem 1.8 in the case $f \equiv f_k$ for no k. There is a posterior distribution for k, computed in (2.13). Fix a large positive integer K. Theories with $k < K$ or $k > n - K$ have negligible posterior mass. For the "mid-zone," theories k with $K \le k \le n - K$, the posterior piles up around the MLE, and the MLE is close to the true parameter.

The assertion about the theory weights has to be proved almost surely as $n \to \infty$, and the predictive probabilities $\rho_{k,n}$ of equations (2.8) and (2.9) have to be estimated. For each k, $\rho_{k,n} \to 0$ a.s. at the rate $\exp\{2^n\kappa + o(2^n)\}$, where κ is an entropy. To make this precise, zones are needed.

ZONE I. $0 \le k < K$, where K is a fixed positive integer.

The posterior weight on theory k is of order $\exp[2^n\kappa + o(2^n)]$, where the entropy $\kappa = \int H(f_k)$ is negative, but increases with k. As the data come in, early theories become less likely than later ones.

THE MID-ZONE. $K \le k \le n - K$. These are the theories that count—as a group. No particular theory survives.

ZONE II. $n - K < k < n$. Fix j. The posterior weight on theory $n - j$ is of order $\exp[2^n\kappa' + o(2^n)]$, where $\kappa' < \int H(f) - \varepsilon_j$ and $\varepsilon_j > 0$. Theory $n - j$ yields to theory l, where l is fixed but large. In the long run, theory l becomes obsolete too, but it stays plausible enough to eliminate theory $n - j$.

ZONE III. $n \le k < \infty$. The total posterior weight on theories $k \ge n$ is of order $\exp[2^n \kappa'' + o(2^n)]$, where the relative entropy $\kappa'' = \int H(f, g_\infty) < \int H(f)$, because $f \ne g_\infty$ by assumption. Again, Zone III bows to theory l.

The posterior piles up around the MLE. For the theories that matter, the posterior $\tilde{\pi}_{k,n}$ piles up around the MLE $\hat{p}_k$, which takes the value $\hat{p}_s = X_s/2^{n-k}$ on $s \in C_k$: See (2.1) or Definition 4.7. (The MLE depends on n and the data, not shown in the notation.) The piling-up has to be established uniformly in k for $1 \le k \le n - K$, almost surely as $n \to \infty$.

The MLE is nearly right. $\|\hat{p}_k - f_k\|_2$ is small uniformly in k for $1 \le k \le n - K$, almost surely as $n \to \infty$. (Alas, $\hat{p}_k - f_k$ will not converge to 0 for $k = n - K$ with K finite, because there are only a finite number of observations on each type of subject.) On the other hand, $\|f_k - f\|_2 \to 0$ as $k \to \infty$. So $\tilde{\pi}_{k,n}$ piles up around f, completing the sketch of proof.

Theory weights, Zone I, $0 \le k < K$. Coming back to rigor, for $x \in C_\infty = \{0,1\}^\infty$ and $s \in C_k = \{0,1\}^k$, let

$$f_k(s) = \int_{C_\infty} f(sx)\lambda^\infty(dx) = E\{f|(x_1, \dots, x_k) = s\}. \tag{4.1}$$

We may extend f_k to C_∞ by setting $f_k(x) = f_k(x_1, \dots, x_k)$. Then

$$f_k(x) = E\{f|x_1, \dots, x_k\}. \tag{4.2}$$

(4.3) LEMMA. *The sequence f_k is a martingale, converging to f a.e. relative to λ^∞ and in L_2, so $\|f_k - f\|_2 \to 0$ as $k \to \infty$.*

(4.4) LEMMA. *The sequence $h_k = \int_{C_\infty} H[f_k(x)]\lambda^\infty(dx)$ is nondecreasing, and converges to $\int_{C_\infty} H[f(x)]\lambda^\infty(dx)$. Furthermore, $h_j < h_k$ for $j < k$ unless $f_j \equiv f_k$.*

Lemma 4.3 is routine. Lemma 4.4 follows from Lemma 4.3 and Jensen's inequality, because the entropy function H in (3.1) is strictly convex.

(4.5) LEMMA. *In a balanced design of order n, the response variables $\eta(t)$ are independent for $t \in C_n$, and $P_f\{\eta(t) = 1\} = f_n(t)$.*

The probability in Lemma 4.5 is unconditional, averaged over the covariates; so is independence; and the proof is routine. P_f is the probability on the sample space Ω that makes the response variables and covariates distributed like balanced designs, according to (1.1). Unless noted otherwise, expectations and variances are relative to P_f.

Recall X_s from (2.1); more explicitly,

$$X_s = \sum_{u \in C_{n-k}} \eta(su). \tag{4.6}$$

(4.7) DEFINITION. Let $\hat{p}_s = X_s/2^{n-k}$. The MLE $\hat{p}_k$ takes the value $\hat{p}_s$ on $s \in C_k$. We extend $\hat{p}_k$ to C_∞ by setting $\hat{p}_k(x) = \hat{p}_k(x_1, \dots, x_k)$.

(4.8) LEMMA. *For a balanced design of order n and $s \in C_k$ with $k < n$, the variables $\hat{p}_s$ are independent, $0 \le \hat{p}_s \le 1$, $E\{\hat{p}_s\} = f_k(s)$, and* $\operatorname{var} \hat{p}_s \le 1/(4 \cdot 2^{n-k})$.

The routine proof is omitted.

Lemmas 4.9 and 4.10 and Corollary 4.11 are first results on the MLE, more specifically, the merging of $\hat{p}_k$ with f_k. Corollary 4.11 will be used in proving Lemma 4.12, which estimates $\rho_{k,n}$.

(4.9) LEMMA. *Fix ε with $0 \le \varepsilon < 1$. For a balanced design of order n and $s \in C_k$ with $k \le n$:*

(a) $P_f\{|\hat{p}_s - f_k(s)| > \varepsilon\} < 1/(4\varepsilon^2 2^{n-k})$.
(b) $P_f\{|\hat{p}_s - f_k(s)| > \varepsilon\} < 2\exp\{-(1/4)\varepsilon^2 2^{n-k}\}$.

PROOF. Claim (a). Use Lemma 4.8 and Chebychev's inequality.

Claim (b). Essentially, this is Bernstein's inequality. To get the precise form of the bound, use (4) in Freedman (1973), noting that $\varepsilon < 1$ and $f_k(s) \le 1$. Also see Gilliland, Hannan and Huang (1976) or Theorem 2 in Hoeffding (1963). □

(4.10) LEMMA. *Choose D so that $D \log 2 > 1$. Fix $\varepsilon > 0$. Almost surely, for all sufficiently large n, in balanced designs of order n, simultaneously for all $s \in C_k$ with $k < n - D\log n$,*

$$|\hat{p}_s - f_k(s)| \le \varepsilon.$$

NOTE. "Almost sure" statements are with respect to P_f.

PROOF. Use Lemma 4.9b and the Borel–Cantelli lemma. The critical sum is bounded above by

$$\sum_{n=1}^{\infty} \sum_{k=0}^{n-D\log n} 2^k \exp\{-\tfrac{1}{4}\varepsilon^2 2^{n-k}\}$$
$$< \sum_{n=1}^{\infty} 2^n \sum_{j=D\log n}^{\infty} 2^{-j} \exp\{-\tfrac{1}{4}\varepsilon^2 2^j\}$$
$$< \infty. \qquad \square$$

(4.11) COROLLARY. *With balanced designs of order n, as $n \to \infty$:*

(a) $\sup_k\{\|\hat{p}_k - f_k\|_\infty\colon 0 \le k < n - D\log n\} \to 0$ *almost surely.*
(b) $\sup_k\{\|\hat{p}_k - f_k\|_2\colon 0 \le k < n - D\log n\} \to 0$ *almost surely.*

(4.12) LEMMA. *For each k, with balanced designs and Γ-uniform π_k, almost surely as $n \to \infty$,*

$$\frac{1}{2^n} \log \rho_{k,n} \to \int_{C_\infty} H[f_k(x)] \lambda^\infty(dx).$$

PROOF. Recall ϕ and ρ from (2.3b and 2.8). Clearly,

$$\frac{1}{2^n} \log \rho_{k,n} = \frac{1}{2^n} \sum_{s \in C_k} \log \phi(2^{n-k}, X_s, \gamma_s). \tag{4.13}$$

Let C_k^0 be the set of $s \in C_k$ with $f_k(s) = 0$. Likewise, $s \in C_k^1$ if and only if $f_k(s) = 1$. And $s \in C_k^+$ if and only if $0 < f_k(s) < 1$. We split the sum in (4.13) into three corresponding parts, and deal with them separately. If $s \in C_k^0$, then $X_s = 0$ a.e. and $\phi(2^{n-k}, X_s, \gamma_s) \leq B/(1 + 2^{n-k})$ by Lemma 3.3e. The contribution to (4.13) from C_k^0 is $o(1)$. Of course, if $s \in C_k^0$, then $H[f_k(s)] = 0$, because $H(0) = 0$. Likewise for C_k^1. For $s \in C_k^+$, we use Lemma 3.3a to estimate ϕ. As $n \to \infty$, the right-hand side of (4.13) is almost surely

$$\frac{1}{2^n} \sum_{s \in C_k} 2^{n-k} H(\hat{p}_s) + o(1) = \frac{1}{2^k} \sum_{s \in C_k} H[f_k(s)] + o(1).$$

Indeed, by Corollary 4.11, $\hat{p}_s$ is close to $f_k(s)$; and H is continuous. Finally,

$$\frac{1}{2^k} \sum_{s \in C_k} H[f_k(s)] = \int_{C_\infty} H[f_k(x)] \lambda^\infty(dx). \qquad \square$$

REMARK. Thus, $\rho_{k,n}$ and $\tilde{w}_{k,n}$ are of order $\exp[\kappa 2^n + o(2^n)]$ where κ depends on k. The idea is that κ increases with k, so posterior mass shifts to higher-order theories as more data comes in. In Lemma 4.12, k is fixed and C_k is finite, so use of Corollary 4.11 is overkill.

(4.14) LEMMA. *Fix $K \geq 0$. Suppose $f \not\equiv f_K$. With balanced designs and Γ-uniform π_k,*

$$\sum_{k=0}^{K} \tilde{w}_{k,n} \Big/ \sum_{k=0}^{\infty} \tilde{w}_{k,n} \to 0 \quad \textit{almost surely as } n \to \infty.$$

PROOF. Fix $k \leq K$. Consider the indices l with $w_l > 0$, so $l \to \infty$ and $f_l \to f$. Find $l > K$ with $w_l > 0$ and $f_l \not\equiv f_K$, so $f_l \not\equiv f_k$ for all $k \leq K < l$. Then $\int H(f_k)\, d\lambda^\infty < \int H(f_l)\, d\lambda^\infty$ by Lemma 4.4. By Lemma 4.12,

$$\lim_{n \to \infty} \frac{1}{2^n} \log \rho_{k,n} < \lim_{n \to \infty} \frac{1}{2^n} \log \rho_{l,n}.$$

By (2.12), $\tilde{w}_{k,n}/\tilde{w}_{l,n} \to 0$. $\square$

Informally, $\tilde{w}_{k,n}$ and $\tilde{w}_{l,n}$ are of order $\exp[\kappa 2^n + o(2^n)]$ and $\exp[\lambda 2^n + o(2^n)]$, respectively; and $\kappa = \int H(f_k)\, d\lambda^\infty < \int H(f_l)\, d\lambda^\infty = \lambda$. So theory k yields to theory l, and Zone I yields to the mid-zone, completing the argument for Zone I.

Theory weights, Zone II, $n - K < k < n$. Fix j with $0 < j < K$. Let $k = n - j$. As in (4.13),

$$\frac{1}{2^n} \log \rho_{n-j,n} = \frac{1}{2^{n-j}} \sum_{s \in C_{n-j}} Z_s, \tag{4.15}$$

where

$$Z_s = \frac{1}{2^j} \log \phi(2^j, X_s, \gamma_s); \tag{4.16}$$

X_s was defined in (4.6) and ϕ in (2.3b).

(4.17) LEMMA. *Fix j with $0 < j < K$. With balanced designs and Γ-uniform π_k, almost surely, as $n \to \infty$,*

$$\frac{1}{2^{n-j}} \left(\sum_{s \in C_{n-j}} Z_s - \sum_{s \in C_{n-j}} E\{Z_s\} \right) \to 0.$$

PROOF. By Definition 1.7 of Γ-uniformity, $\gamma_s \geq b > 0$ and then by Lemma 3.3d the Z_s are bounded between $-G$ and 0, where $G = 1 + |\log b|/2$. So $|Z_s - E\{Z_s\}| < G$ and $\operatorname{var}\{Z_s\} < G^2$. Furthermore, the Z_s are independent as s ranges over C_k, by Lemma 4.8. By Chebychev's inequality, for $\delta > 0$,

$$P_f \left\{ \sum_{s \in C_{n-j}} Z_s - \sum_{s \in C_{n-j}} E\{Z_s\} > \delta 2^{n-j} \right\} < G^2/(\delta^2 2^{n-j}),$$

which sums in n for each fixed j. The Borel–Cantelli lemma completes the proof. □

(4.18) LEMMA. *Fix $j = 1, 2, \dots$. Let $m = 2^j$. Recall $\varepsilon_m > 0$ from Lemma* 3.8b. *With balanced designs and Γ-uniform π_k, almost surely,*

$$\limsup_{n \to \infty} \frac{1}{2^n} \log \rho_{n-j,n} \leq \left(\int_{C_\infty} H[f(x)] \lambda^\infty(dx) \right) - \varepsilon_m.$$

PROOF. Recall the definition of Z_s from (4.16). By Lemmas 4.5, 4.8 and 3.7,

$$E\{Z_s\} \leq \psi(2^j, f_{n-j}(s), \gamma_s). \tag{4.19}$$

For $\delta > 0$, and n sufficiently large,

$$\begin{aligned}
\frac{1}{2^n}\log\rho_{n-j,n} &= \frac{1}{2^{n-j}}\sum_{s\in C_{n-j}} Z_s \quad \text{by (4.15)}\\
&\le \frac{1}{2^{n-j}}\sum_{s\in C_{n-j}} E\{Z_s\} + \delta \quad \text{by Lemma 4.17}\\
&\le \frac{1}{2^{n-j}}\sum_{s\in C_{n-j}} \psi\big(2^j, f_{n-j}(s), \gamma_s\big) + \delta \quad \text{by (4.19)}\\
&\le \int_{C_\infty} H\big[f_{n-j}(x)\big]\lambda^\infty(dx) - \varepsilon_m + \delta \quad \text{by Lemma 3.8b}\\
&\to \int_{C_\infty} H[f(x)]\lambda^\infty(dx) - \varepsilon_m + \delta \quad \text{by Lemma 4.4.}
\end{aligned}$$

This proves Lemma 4.18, since δ was arbitrary. □

(4.20) LEMMA. *Fix $K > 0$. With balanced designs and Γ-uniform π_k,*

$$\sum_{k=n-K}^{n-1}\tilde{w}_{k,n}\Big/\sum_{k=0}^{\infty}\tilde{w}_{k,n} \to 0 \quad \textit{almost surely as } n\to\infty.$$

PROOF. Fix j with $1 \le j \le K$. By Lemma 4.18, almost surely, for all sufficiently large n,

$$\frac{1}{2^n}\log\rho_{n-j,n} < \int H(f)\,d\lambda^\infty - \varepsilon_m/2. \tag{4.21}$$

Using Lemma 4.4, find l large with $w_l > 0$ and

$$\int H(f_l)\,d\lambda^\infty > \int H(f)\,d\lambda^\infty - \varepsilon_m/4. \tag{4.22}$$

Combine (4.21) and (4.22): almost surely, for all sufficiently large n,

$$\frac{1}{2^n}\log\rho_{n-j,n} \le \int H(f_l)\,d\lambda^\infty - \varepsilon_m/4. \tag{4.23}$$

By Lemma 4.12, almost surely, for all sufficiently large n,

$$\frac{1}{2^n}\log\rho_{l,n} \ge \int H(f_l)\,d\lambda^\infty - \varepsilon_m/8. \tag{4.24}$$

Combine (4.23) and (4.24):

$$\tilde{w}_{n-j,n}/\tilde{w}_{l,n} \to 0.$$ □

Informally, $\tilde{w}_{n-j,n}$ and $\tilde{w}_{l,n}$ are of order $\exp[\kappa 2^n + o(2^n)]$ and $\exp[\lambda 2^n + o(2^n)]$, respectively; and $\kappa < \lambda$, because κ is further below $\int H(f)\,d\lambda^\infty$. Thus, theory $n-j$ yields to theory l, and Zone II yields to the mid-zone.

Theory weights, Zone III, $n \le k < \infty$. Abbreviate $L_{k,n} = \log \rho_{k,n}$. The relative entropy function $H(\cdot, \cdot)$ was defined in (2.10). Write $g_\infty(t)$ for the common value of $g_\infty(x)$ over $x \in C_\infty$ with $x_j = t_j$ for $1 \le j \le n$. By Lemma 2.11, for all sufficiently large n,

$$L_{k,n} = \sum_{t \in C_n} H[\eta_t, g_\infty(t)] \quad \text{for all } k \ge n. \tag{4.25}$$

In particular, in this range $L_{k,n}$ does not depend on k. Let

$$h_n^* = \sum_{t \in C_n} H[f_n(t), g_\infty(t)] = 2^n \int_{C_\infty} H(f_n, g_\infty)\, d\lambda^\infty \tag{4.26}$$

and

$$\begin{aligned} T_n &= \sum_{t \in C_n} \left[H[\eta_t, g_\infty(t)] - H[f_n(t), g_\infty(t)] \right] \\ &= \sum_{t \in C_n} \left[[\eta_t - f_n(t)] \log \frac{g_\infty(t)}{1 - g_\infty(t)} \right], \end{aligned} \tag{4.27}$$

where g_∞ is bounded away from 0 and 1 by Definition 1.7 of Γ-uniformity. Clearly,

$$L_{k,n} = h_n^* + T_n \quad \text{for all } k \ge n. \tag{4.28}$$

(4.29) LEMMA. *With balanced designs and Γ-uniform π_k, almost surely, for all sufficiently large n, $|T_n| \le \sqrt{2^n}\, n$.*

PROOF. Use Chebychev's inequality and the Borel–Cantelli lemma. □

(4.30) LEMMA. *Suppose $f \ne g_\infty$. With balanced designs and Γ-uniform π_k, almost surely as $n \to \infty$,*

$$\sum_{k=n}^{\infty} \tilde{w}_{k,n} \bigg/ \sum_{k=0}^{\infty} \tilde{w}_{k,n} \to 0$$

PROOF. Since $f \ne g_\infty$, $\int H(f)\, d\lambda^\infty - \int H(f, g_\infty)\, d\lambda^\infty = \varepsilon > 0$. Now

$$\begin{aligned} \frac{1}{2^n} h_n^* &= \int H(f_n, g_\infty)\, d\lambda^\infty \quad \text{by (4.26)} \\ &< \int H(f, g_\infty)\, d\lambda^\infty + \varepsilon/2 \\ &< \int H(f)\, d\lambda^\infty - \varepsilon/2. \end{aligned}$$

The second line holds for all sufficiently large n, because $f_n \to f$. Furthermore, $T_n/2^n < \varepsilon/4$ for all sufficiently large n, by Lemma 4.29. By (4.28) and (2.12),

$$\sum_{k=n}^{\infty} \tilde{w}_{k,n} = \left(\sum_{k=n}^{\infty} w_k \right) \exp(h_n^* + T_n)$$
$$< \left(\sum_{k=0}^{\infty} w_k \right) \exp\left(2^n \left[\int H(f)\, d\lambda^\infty - \varepsilon/4 \right] \right).$$

As in Lemma 4.20, choose l with $w_l > 0$ and $\int H(f_l)\, d\lambda^\infty > \int H(f)\, d\lambda^\infty - \varepsilon/8$. By Lemma 4.12, almost surely, for all sufficiently large n,

$$\tilde{w}_l > w_l \exp\left(2^n \left[\int H(f)\, d\lambda^\infty - \varepsilon/8 \right] \right).$$

Thus, theories $n, n+1, \ldots$ have negligible posterior weight, by comparison with theory l. Here, $n \to \infty$ while l is fixed but large. □

Informally, theories $n, n+1, \ldots$ have total posterior weight of order $\exp[\kappa 2^n + o(2^n)]$; theory l has posterior weight of order $\exp[\lambda 2^n + o(2^n)]$; $\lambda \doteq \int H(f)\, d\lambda^\infty$ for l large, but $\kappa = \int H(f, g_\infty)\, d\lambda^\infty < \int H(f)\, d\lambda^\infty$. All theories in Zone III, combined, yield to theory l. So Zone III is a posteriori dominated by the mid-zone, completing the argument for Zone III.

REMARK. If $f \equiv g_\infty$, so does f_k, and $\int H(f_k)$ will be constant for most theories k—except for a few near 0 and a few just below n. This is a more delicate case, to be considered in the next section. So far, it has only been necessary to estimate $\rho_{k,n}$ for one k at a time; in the next section, uniform estimates will be needed for ranges of k's.

The posterior piles up around the MLE. Fix a nonnegative integer k, and small positive numbers δ and ε. Define $G \subset \Theta_k \times \Omega$ as follows: $(f, \omega) \in G$ if and only if $f \in \Theta_k$ and $|\theta_s - \hat{p}_s(\omega)| \le \varepsilon$ for all but at most $\delta 2^k$ strings $s \in C_k$. The set G depends on δ, ε, k and n.

(4.31) PROPOSITION. *Fix δ, ε, $\delta' > 0$. Suppose the π_k are Γ-uniform, and the designs are balanced. There is a $K < \infty$ such that $\tilde{\pi}_{k,n}\{G\} > 1 - \delta'$ for all $\omega \in \Omega$ and all n, k with $K \le k \le n - K$.*

PROOF. Corollary 2.6 in Diaconis and Freedman (1990) establishes that for some $\psi(\varepsilon) > 0$, for all $s \in C_k$,

$$\tilde{\pi}_{k,n}\{|\theta_s - \hat{p}_s| > \varepsilon\} \le 1\big/\left[1 + \psi(\varepsilon)\exp(2 \cdot 2^{n-k} \cdot \varepsilon^2)\right]$$
$$\le \delta/2 \quad \text{for } 0 \le k \le n - K, \tag{4.32}$$

provided K Is large enough.

From the point of view of $\tilde{\pi}_{k,n}$, the events $|\theta_s - \hat{p}_s| > \varepsilon$ are independent as s ranges over C_k, by Lemma 2.4; each event has probability at most $\delta/2$, by

(4.32). The $\tilde{\pi}_{k,n}$-chance that more than $\delta 2^k$ of these events occur can be estimated by Chebychev's inequality:

$$1 - \tilde{\pi}_{k,n}\{G\} < 4/(\delta^2 2^k) < \delta'$$

for all n and k with $K \le k \le n - K$, provided K is large enough. □

The basic neighborhoods $N(f, \delta, \varepsilon)$ were given in Defintion 1.4, and the MLE $\hat{p}_k$ in Definition 4.7.

(4.33) COROLLARY. *Fix δ, ε, $\delta' > 0$. Suppose the π_k are Γ-uniform and the designs are balanced. There is a $K < \infty$ such that $\tilde{\pi}_{k,n}\{N(\hat{p}_k, \delta, \varepsilon)\} > 1 - \delta'$ for all $\omega \in \Omega$ and n, k with $K \le k \le n - K$.*

REMARK. Although Corollary 2.6 in Diaconis and Freedman (1990) is correct, there is a minor error in the proof: ε_h should be defined as $[g(h) - 2h^2]/g(h)$, not $g(h) - 2h^2$. The h there corresponds to ε in Proposition 4.31. The $\psi(h)$ is not related to $\psi(m, p)$, but is positive. It is remarkable that Proposition 4.31 holds for all ω: There is no exceptional null set to eliminate.

If $f \equiv f_k$ for some k, theory k counts; and Corollary 4.33 is not enough. The next proposition covers theories in the range $0 \le k \le n - D \log n$, and modifies the definition of G. Fix $\varepsilon > 0$. Let $(f, \omega) \in G$ if and only if $f \in \Theta_k$ and $|\theta_s - \hat{p}_s(\omega)| \le \varepsilon$ for all $s \in C_k$.

(4.34) PROPOSITION. *Fix ε, $\delta' > 0$. Choose $D < \infty$ so $D \log 2 > 1$. Suppose π_k is Γ-uniform, and the designs are balanced. There is a finite $n_0 = n_0(\varepsilon, \delta', D)$ such that $\tilde{\pi}_{k,n}\{G\} > 1 - \delta'$ for all $\omega \in \Omega$ and all n, k with $0 \le k \le n - D \log n$, provided $n > n_0$.*

The proof of Proposition 4.34 is like that of Proposition 4.31, but using the Bonferroni inequality:

$$1 - \tilde{\pi}_{k,n}\{G\} \le 2^k \big/ \big[1 + \psi(\varepsilon)\exp(2 \cdot 2^{n-k} \cdot \varepsilon^2)\big] \to 0$$

as $n \to \infty$, uniformly for $k \le n - D \log n$. The range of k's covered by Proposition 4.34 overlaps that of Proposition 4.31; however, Proposition 4.34 covers k's near 0 while Proposition 4.31 gets a little closer to n. For the k's they both cover, Proposition 4.34 is better.

For all k with $0 \le k \le n - D \log n$, $\hat{p}_s$ stays close to $f_k(s)$ for all $s \in C_k$, as in Lemma 4.10. The proof of Proposition 4.34 uses the condition $k \le n - D \log n$ from another perspective, to make the bound on $1 - \tilde{\pi}_{k,n}\{G\}$ go to 0.

(4.35) COROLLARY. *Fix δ, ε, $\delta' > 0$. Suppose the π_k are Γ-uniform and the designs are balanced. There is a finite $n_0 = n_0(\varepsilon, \delta', D)$ such that $\tilde{\pi}_{k,n}\{N(\hat{p}_k, \delta, \epsilon)\} > 1 - \delta'$ for all $\omega \in \Omega$ and all n, k with $0 \le k \le n - D \log n$, provided $n > n_0$.*

(4.36) COROLLARY. *Fix δ, ε, $\delta' > 0$. Suppose the π_k are Γ-uniform and the designs are balanced. There is a $K < \infty$ such that $\tilde{\pi}_{k,n}\{N(\hat{p}_k, \delta, \varepsilon)\} > 1 - \delta'$ for all k with $0 \le k \le n - K$.*

This is immediate from Corollaries 4.33 and 4.35.

The MLE is nearly right. Corollary 4.11 establishes merging of $\hat{p}_k$ with f_k for $0 \le k < n - D\log n$, where $D\log 2 > 1$. The next result establishes it for $D\log n \le k \le n - K$, the lower end of the range being redundant. Recall the empirical probabilities $\hat{p}_s$ from Definition 4.7, and $f_k(s)$ from (4.1).

(4.37) PROPOSITION. *Fix δ, $\varepsilon > 0$. Suppose the designs are balanced. There is a positive, finite K such that, almost surely, for all sufficiently large n, for all k with $D\log n \le k \le n - K$, for fewer than $\delta 2^k$ strings $s \in C_k$,*

$$|\hat{p}_s - f_k(s)| > \varepsilon. \tag{4.38}$$

PROOF. By Lemma 4.8, the events defined by (4.38) are independent as s varies over C_k. By Lemma 4.9a, each event has probability less than

$$1/(4 \cdot 2^{n-k} \cdot \varepsilon^2) < \delta/2$$

provided $k \le n - K$ and K is sufficiently large. The chance that $\delta 2^k$ or more of these events occur is at most $4/(\delta^2 2^k)$, by Chebychev's inequality.

We must show

$$\sum_n \sum_{k=a_n}^{b_n} 1/2^k < \infty. \tag{4.39}$$

The lower limit on the inner sum is $a_n = D\log n$; the upper limit is $b_n = n - K$. The inner sum is of order $1/2^{a_n} = O(1/n^{D\log 2})$. This proves (4.39), and the Borel–Cantelli lemma completes the argument. □

NOTE. Proposition 4.37 involves the "objective" probability P_f on the sample space Ω, while Proposition 4.31 involves the "subjective" $\tilde{\pi}_{k,n}$ on the parameter space Θ. However, the proofs are virtually the same.

(4.40) LEMMA. *Fix $\varepsilon > 0$. Suppose the designs are balanced. There is a $K < \infty$ such that $\|\hat{p}_k - f_k\|_2 < \varepsilon$ for all k with $0 \le k \le n - K$, almost surely, for all sufficiently large n.*

PROOF. This is immediate from Lemma 4.10 and Proposition 4.37. □

THE PROOF OF THEOREM 1.8. Combining Corollary 4.36 and Lemma 4.40 gives Lemma 4.41a; δ and ε in Corollary 4.36 and Lemma 4.40 must be computed from the δ and ε in Lemma 4.41. Then use Lemma 4.3 to get Lemma 4.41b.

(4.41) LEMMA. *Fix δ, ε, $\delta' > 0$. Suppose the π_k are Γ-uniform and the designs are balanced. There is a $K < \infty$ such that almost surely, for all sufficiently large n:*

(a) $\tilde{\pi}_{k,n}\{N(f_k, \delta, \varepsilon)\} > 1 - \delta'$ *for all k with* $0 \le k \le n - K$.
(b) $\tilde{\pi}_{k,n}\{N(f, \delta, \varepsilon)\} > 1 - \delta'$ *for all k with* $K \le k \le n - K$.

Suppose $f \equiv f_k$ for no k. Theorem 1.8 will now be proved under a side-condition, that $f \equiv f_k$ for no k. Recall that $\|\mu - \nu\|$ is the variation distance between μ, $\nu \in \Pr(\Theta)$; the posterior $\tilde{\pi}_n$ was computed in Lemma 2.14. Fix a large, finite K. Let

$$\tilde{\pi}_n^K = \sum_{k=K}^{n-K} \tilde{w}_{k,n}\tilde{\pi}_{k,n} \Big/ \sum_{k=K}^{n-K} \tilde{w}_{k,n}.$$

Combine Lemmas 4.14, 4.20, 4.30 and 3.10 to see that

$$\|\tilde{\pi}_n - \tilde{\pi}_n^K\| \to 0 \quad \text{as } n \to \infty, \text{ almost surely.}$$

Fix δ, ε, $\delta' > 0$ and use Lemma 4.41b: almost surely, for all sufficiently large n, for all k with $K \le k \le n - K$,

$$\tilde{\pi}_{k,n}\{N(f, \delta, \varepsilon)\} > 1 - \delta'.$$

So,

$$\lim_{n\to\infty} \tilde{\pi}_n\{N(f, \delta, \varepsilon)\} = 1 \quad \text{almost surely.}$$

This completes the proof of Theorem 1.8, provided $f \equiv f_k$ for no k.

Suppose $f \equiv f_K$ for some K, but $f \not\equiv g_\infty$. The remaining case in the proof of Theorem 1.8 can now be handled: Suppose $f \equiv f_K$ for some K, but $f \not\equiv g_\infty$. Lemma 4.14 shows that all theories $k < K$ can be ignored. Lemmas 4.20 and 4.30 eliminate $k > n - K$. Suppose $K \le k \le n - K$. Then $f_k \equiv f$, and the argument proceeds from Lemma 4.41a rather than Lemma 4.41b. This completes the proof of Theorem 1.8. □

5. Proof of Theorem 1.9. The proof of Theorem 1.9 is rather like that of Theorem 1.8, but now $g_\infty \equiv f$. In other words, the mean of the prior happens to be exactly equal to the true f. Oddly, that is the delicate case. Indeed, $\rho_{k,n}$ turns out to be of order

$$\exp\left[2^n \int H(f) - \beta(n-k)2^k + o[(n-k)2^k]\right]$$

for all k except those just less than n. Here, $\beta = (1/2)\log 2$. We must estimate $L_{k,n} = \log \rho_{k,n}$ to order $(n-k)2^k$ or better, and uniformly in k. The factor $1/\sqrt{m}$ in (3.2) provides crucial leverage.

In Theorem 1.9, there are two cases, according as g_∞ is constant or not. Only the first case will be done, where the analysis is a bit easier. The second case can be handled by splitting Θ and then using similar arguments: indeed, g_∞ is piecewise constant on Θ by Definition 1.7.

We are assuming that for some $p \in (0, 1)$,

$$f(x) \equiv g_\infty(x) \equiv p \quad \text{for all } x \in C_\infty. \tag{5.1}$$

Recall Definition 1.7 of Γ-uniformity. The index n_1 was defined in Definition 1.7, and $g_k \equiv g_\infty$ for $k \geq n_1$. By Definition 1.7 and (4.1),

$$f_k(x) \equiv p \quad \text{for all } x \in C_\infty \text{ and all } k \geq 0 \tag{5.2a}$$

and

$$g_n(x) \equiv p \quad \text{for all } x \in C_\infty \tag{5.2b}$$

provided (as we will assume throughout)

$$n \geq n_1. \tag{5.2c}$$

Results on $L_{k,n}$ are summarized in Proposition 5.5. To motivate the form of the results, consider $L_{k,n}$ for $k \geq n$. By (4.26)–(4.28) and a bit of algebra based on (5.1),

$$L_{k,n} = 2^n H(p) + T_n \quad \text{for } k \geq n, \tag{5.3a}$$

where

$$T_n = H'(p) \sum_{t \in C_n} (\eta_t - p). \tag{5.3b}$$

The random term T_n is of order $\sqrt{2^n}$, and turns up (somewhat surprisingly) in all the $L_{k,n}$, even for $k < n$. Therefore, T_n does not affect likelihood ratios—and further terms in asymptotic expansions are needed. For $0 \leq k \leq n - D \log n$, we can approximate $L_{k,n}$ by

$$\alpha_{k,n} = 2^n H(p) + T_n - \beta(n-k)2^k. \tag{5.4a}$$

For late k's, there is an additional term. To define it, let N_k be the number of $s \in C_k$ with $X_s = 0$ or 2^{n-k}. (For $k \leq n - D \log n$, $N_k = 0$ almost surely; but for k near n, N_k may be appreciable.) Let $s \in D_k$ if and only if $0 < X_s < 2^{n-k}$, and let

$$\Xi_{k,n} = -\beta(n-k)N_k + \sum_{s \in D_k} \log \sqrt{\hat{p}_s(1-\hat{p}_s)}\,. \tag{5.4b}$$

For $n - D \log n \leq k \leq n - 1$, we approximate $L_{k,n}$ by $\alpha_{k,\mathrm{n}} + \Xi_{k,n}$; all terms in $\Xi_{k,n}$ are negative, because $0 \leq \hat{p}_s \leq 1$.

Assume the designs are balanced and the π_k are Γ-uniform; (5.1)–(5.4) are in force. Let ε be small and positive. Choose D with $D \log 2 > 1$. For Proposition 5.5e, choose $K = K(\varepsilon)$ large but finite. Then, for Proposition 5.5f, choose $\varepsilon' = \varepsilon'(K)$ small but positive. Claims 5.5a–f hold uniformly in the indicated range, for all sufficiently large n, almost surely. Write $C_0, C_1, \ldots$ for positive, finite constants, whose exact values do not matter. These constants are distinguished by context from the sets of strings C_k.

(5.5) PROPOSITION. *Assume the conditions of Theorem* 1.9 *and* (5.1)–(5.4). *Fix small positive* ε, ε' *and a large positive integer* K. *Almost surely, for all sufficiently large* n:

(a) *For* $0 \le k \le n - D\log n$, $|L_{k,n} - \alpha_{k,n}| < (C_1 + \varepsilon n)2^k$.
(b) *For* $D\log n \le k \le n - D\log n$, $|L_{k,n} - \alpha_{k,n}| < C_2 2^k$.
(c) *For* $n - D\log n \le k \le n-1$, $|L_{k,n} - \alpha_{k,n} - \Xi_{k,n}| < C_3 2^k$.
(d) *For* $0 \le k \le n - D\log n$, $|L_{k,n} - \alpha_{k,n}| < 2\varepsilon(n-k)2^k$.
(e) *For* $n - D\log n \le k \le n - K$, $|L_{k,n} - \alpha_{k,n} - \Xi_{k,n}| < 2\varepsilon(n-k)2^k$.
(f) *For* $n - K \le k \le n-1$, $L_{k,n} \le 2^n[H(p) - \varepsilon']$.
(g) *For* $k \ge n$, $L_{k,n} = 2^n H(p) + T_n$.

NOTE. For $k < (1/2)n$, T_n matters. For $k > (1/2 + \delta)n$, T_n can be absorbed into the error terms in (b) and (c).

Some preliminary estimates are needed, before proving Proposition 5.5.

(5.6) LEMMA. *Suppose* $n \ge k$ *and* $n \ge n_1$. *As* s *varies over* C_k, *the variables* X_s *are independent and* $\text{bin}(2^{n-k}, p)$.

PROOF. In view of (5.1), this follows from Lemma 4.5 and (4.6). □

Recall the entropy function H from (3.1) and the bound $H^*(p)$ from Definition 3.13. Recall that $\hat{p}_s = X_s/2^{n-k}$. By Lemma 5.6 and Corollary 3.15,

$$H(p) < E\{H(\hat{p}_s)\} < H(p) + H^*(p)p(1-p)/2^{n-k}. \tag{5.7}$$

To help estimate $L_{k,n}$, let

$$Q_{k,n} = 2^{n-k} \sum_{s \in C_k} (\hat{p}_s - p)^2. \tag{5.8}$$

(5.9) LEMMA. *Suppose* $n > k$ *and* $n \ge n_1$.

(a) $E\{Q_{k,n}\} = 2^k p(1-p)$.
(b) *Let* $q = 1 - p$. *Then*

$$\text{var}\{Q_{k,n}\} = 2^k 2p^2q^2\left[1 + pq(1-6pq)/2^{n-k}\right].$$

(c) $15 \cdot 2^k p^2q^2/8 < \text{var}\, Q_{k,n} < 2^k/6$.

PROOF. Claim (a) is elementary, starting from Lemma 5.6; and so is (b) although the algebra is irritating; see Cramer [(1957), page 195]. Claim (c) follows from (b). Indeed, $0 < pq \le 1/4$. The function $x \to x(1-6x)$ on $[0, 1/4]$ is bounded between $1/24$ and $-1/8$. □

(5.10) LEMMA. *Almost surely, for all sufficiently large* n, *for all* k *with* $D\log n \le k \le n-1$, $Q_{k,n} \le 2^k$.

PROOF. By Chebychev's inequality and Lemma 5.9,

$$P\{Q_{k,n} > 2^k\} < \text{const.}/2^k.$$

But

$$\sum_{n=1}^{\infty} \sum_{k=D\log n}^{n-1} 1/2^k < \infty$$

because $D \log 2 > 1$. Now use the Borel–Cantelli lemma. □

A similar argument proves the next result.

(5.11) LEMMA. *Fix $\delta > 0$. Almost surely, for all sufficiently large n, for all k with $0 \le k \le n - 1$, $Q_{k,n} < [p(1-p) + \delta n]2^k$.*

Recall T_n from (5.3). Let

$$R_{k,n} = 2^{n-k} \sum_{s \in C_k} H(\hat{p}_s). \tag{5.12}$$

(5.13) LEMMA. *For all n, all k with $0 \le k \le n-1$, and all $\omega \in \Omega$, with $H^*(p)$ as defined in Definition 3.13,*

$$|R_{k,n} - 2^n H(p) - T_n| \le H^*(p) Q_{k,n}.$$

PROOF. Expand H around p, using (3.11)–(3.14). After a bit of algebra,

$$\left| R_{k,n} - 2^n H(p) - 2^{n-k} H'(p) \sum_{s \in C_k} (\hat{p}_s - p) \right| \le H^*(p) Q_{k,n}.$$

The linear term can be reorganized:

$$2^{n-k} H'(p) \sum_{s \in C_k} (\hat{p}_s - p) = H'(p) \sum_{t \in C_n} (\eta_t - p) = T_n. \tag{5.14}$$ □

(5.15) LEMMA. *Fix $\delta > 0$. Almost surely, for all sufficiently large n, for all $k \le n - D \log n$ and all $s \in C_k$, $|\hat{p}_s - p| \le \delta$.*

PROOF. This is a special case of Lemma 4.10. □

(5.16) LEMMA. *Almost surely, for all sufficiently large n, for all k with $0 \le k \le n - D \log n$,*

$$|L_{k,n} - R_{k,n} + \beta(n-k)2^k| < \text{const. } 2^k.$$

PROOF. $L_{k,n} = \log \rho_{k,n} = \sum_{s \in C_k} \log \phi(2^{n-k}, X_s, \gamma_s)$ by (2.8). Now use Lemma 3.3a to estimate ϕ. Informally, the term $\beta(n-k)2^k$ comes from the $1/\sqrt{m}$ in (3.2). Because $m = 2^{n-k}$,

$$\log(1/\sqrt{m}) = -(\tfrac{1}{2}\log 2)(n-k) = -\beta(n-k).$$

Summing over C_k, with 2^k terms, yields $-\beta(n-k)2^k$. The constant terms in (3.2) and (3.3a), namely,

$$\log\sqrt{2\pi}\,,\qquad \log\sqrt{\hat{p}_s(1-\hat{p}_s)}\,,\qquad \log a\,,\quad \log A,$$

add up to the error term $O(2^k)$. We are using Lemma 5.15 to keep $\hat{p}_s$ bounded away from 0 and 1. □

PROOF OF PROPOSITION 5.5a. Combine Lemmas 5.11, 5.13 and 5.16. □

PROOF OF PROPOSITION 5.5b. Replace Lemma 5.11 by Lemma 5.10 in the above. □

PROOF OF PROPOSITION 5.5c. This is like (5.5b), and the proof of Lemma 5.16. Each $s \in C_k$ with $X_s = 0$ or 2^{n-k} contributes an extra term $-\beta(n-k)$ to $L_{n,k}$, because (3.2) defines $\phi^*(m,j)$ differently for $j = 0$, m and $0 < j < m$. Furthermore, terms involving $\log\sqrt{\hat{p}_s(1-\hat{p}_s)}$ have to be entered explicitly, because $\hat{p}_s$ can be close to 0 or 1. □

PROOF OF PROPOSITION 5.5d. This is immediate from parts (a) – (b); use (a) for theories $k \le D\log n$, and (b) for the range $D\log n \le k \le n - D\log n$. □

PROOF OF PROPOSITION 5.5e. This is immediate from Proposition 5.5c. □

PROOF OF PROPOSITION 5.5f. This follows from Lemma 4.18, because $H(f) \equiv H(p)$ by (5.1). □

PROOF OF PROPOSITION 5.5g. This is just (5.3). □

This completes the proof of Proposition 5.5. The next lemma will help with the consistency argument, and the elementary proof is omitted.

(5.17) LEMMA. *Let* $n \ge 1$.

(a) $x \to \log(n-x) + x\log 2$ *is strictly concave on* $[0, n-1]$, *and strictly increasing on* $[0, n-2]$.

(b) $(n-k)2^k \ge n$ *for* $k = 0, \ldots, n-1$.

The next lemma will help with the inconsistency argument. Recall the basic neighborhood $N(f, \delta, \varepsilon)$ from Definition 1.4; δ and ε are not related to those in Proposition 5.5.

(5.18) LEMMA. *Suppose* $f \equiv p$ *and* π_k *is* Γ*-uniform. Fix* $\delta \in (0, 1/4)$. *There is a small positive* ε *(which does not depend on* δ*) such that* $\tilde{\pi}_{k,n}\{N(f,\delta,\varepsilon)\} \le 1/2^n$ *uniformly in* $k \ge n$ *and* ω.

PROOF. Recall that Θ_k is the set of functions that depend only on the first k bits. The success probabilities $\theta_s(h)$ were defined for $h \in \Theta_k$ by Definition 1.6. If $h \in \Theta_k$, then $h \in N(f, \epsilon, \delta)$ if and only if $|\theta_s(h) - p| \le \epsilon$ for all but $\delta 2^k$ strings $s \in C_k$.

Of course, $\tilde{\pi}_{k,n}(\Theta_k) = 1$. By Lemma 2.7, from the perspective of $\tilde{\pi}_{k,n}$, the random variables θ_s are independent as s ranges over C_k, and θ_s has one of the following densities:

$$\gamma_s, \gamma_{\tau_t}(1, 0, \cdot\,), \qquad \gamma_{\tau_t}(1, 1, \cdot\,).$$

The latter two densities are defined in (2.3), and the normalizing constant $\phi(1, j, \gamma)$ in the denominator is estimated next. By (2.3b), $\phi(1, 1, \gamma) = \int \theta\gamma(\theta)\, d\theta \ge b \int \theta\, d\theta = b/2$, and likewise for $\phi(1, 0, \gamma)$. So each of the three densities is bounded above by $G = 2B/b$. Therefore,

$$\tilde{\pi}_{k,n}\{|\theta_s - p| \le \varepsilon\} \le G\varepsilon \quad \text{for each } s \in C_k.$$

[The constants b and B appear in Definition 1.7 of Γ-uniformity.]

Choose $\varepsilon > 0$ so small that $G\varepsilon < 1/4$. Then $\tilde{\pi}_{k,n}\{N(f, \delta, \varepsilon)\}$ is bounded above by the chance that, among 2^k independent events of probability $G\varepsilon \le 1/4$, at least $(1 - \delta)2^k \ge (3/4)2^k$ will occur. Chebychev's inequality gives the bound $1/2^k \le 1/2^n$. □

PROOF OF THEOREM 1.9.

CLAIM (a). By (2.12), (5.4a) and Proposition 5.5d, for large n,

$$\tilde{w}_{l,n} > w_l \exp\left[2^n H(p) + T_n - \beta(n - l)2^l - 2\varepsilon(n - l)2^l\right]. \tag{5.19}$$

By Proposition 5.5f, theories with $n - K \le k \le n - 1$ have negligible posterior weight. By Proposition 5.5g, theories with $k \ge n$ have posterior weight

$$\begin{aligned}\sum_{k=n}^{\infty} \tilde{w}_{k,n} &= \left(\sum_{k=n}^{\infty} w_k\right) \exp(2^n H(p) + T_n) \\ &< \exp\left(2^n H(p) + T_n - \beta n 2^l - \delta n 2^l\right) \quad \text{for } n \text{ large.}\end{aligned} \tag{5.20}$$

Now use the fact that ε in Proposition 5.5d is arbitrary: Choose it so small that $2\varepsilon < \delta$. Compare (5.19) and (5.20) to see that theories with $k \ge n$ are negligible. Indeed, $\Sigma_{k=n}^{\infty} \tilde{w}_{k,n} \ll \tilde{w}_{l,n}$ because $-\delta 2^l < -2\varepsilon 2^l$. The index l is fixed and the term $(\beta + 2\varepsilon)l2^l$ does not affect the reasoning. Posterior weight concentrates on theories with $k \le n - K$, and $\tilde{\pi}_{k,n}$ piles up around $f_k \equiv f$, by Lemma 4.41a.

CLAIM (b). This is the reverse side of (a). Consider only n with $\Sigma_{k=n}^{\infty} w_k > \exp(-\beta n 2^l + \delta n 2^l)$. As in (5.20),

$$\begin{aligned}\sum_{k=n}^{\infty} \tilde{w}_{k,n} &= \left(\sum_{k=n}^{\infty} w_k\right) \exp\left[2^n H(p) + T_n\right] \\ &> \exp\left[2^n H(p) + T_n - \beta n 2^l + \delta n 2^l\right].\end{aligned} \tag{5.21}$$

Fix $K > l + 2$. For theories with $k \leq n - K$, by Proposition 5.5d, 5.5e and Lemma 5.17b,

$$\begin{aligned}\sum_{k=0}^{n-K} \tilde{w}_{k,n} &< \sum_{k=l}^{n-K} w_k \exp\left[2^n H(p) + T_n - \beta(n-k)2^k + 2\varepsilon(n-k)2^k\right] \\ &< \left(\sum_{k=l}^{\infty} w_k\right) \exp\left[2^n H(p) + T_n - \beta(n-l)2^l + 2\varepsilon(n-l)2^l\right].\end{aligned} \tag{5.22}$$

Since $\Xi_{k,n} < 0$, it was dropped on the right-hand side of (5.22): see (5.4b). Compare (5.21) and (5.22): $\sum_{k=0}^{n-K} \tilde{w}_{k,n} \ll \sum_{k=n}^{\infty} \tilde{w}_{k,n}$. Theories with $n - K \leq k \leq n - 1$ are also negligible. It is theories with $k \geq n$ which dominate, and $\tilde{\pi}_n$ is close in variation distance to

$$\sum_{k=n}^{\infty} w_n \tilde{\pi}_{k,n} \Big/ \sum_{k=n}^{\infty} w_k,$$

by Lemma 3.10. Lemma 5.18 completes the proof: The posterior mass in a basic neighborhood of f tends to 0. □

Acknowledgment. We would like to thank two very helpful referees.

REFERENCES

BOWMAN, A. (1984). An alternative method of cross-validation for the smoothing of densities. *Biometrika* **71** 353–360.

BREIMAN, L. AND FREEDMAN, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

BRUNO, A. (1964). Sugli eventi pazialmente scambiabili. *Giornale dell'Instituto Italiano degli Attuari* **27**. [English translation in de Finetti (1972)].

COX, D. (1988). Approximation of least squares regression on nested subspaces. *Ann. Statist.* **16** 713–732.

COX, D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1693.

CRAMER, H. (1957). *Mathematical Methods of Statistics*. Princeton Univ. Press.

DATTA, S. (1991). On the consistency of posterior mixtures and its applications. *Ann. Statist.* **19** 338–353.

DE FINETTI, B. (1959). La probabilità, la statistica, nei rapporti con l'induzione, secondo diversi punti di vista. *Centro Internazionale Matematica Estivo Cremonese, Rome*. [English translation in de Finetti (1972)].

DE FINETTI, B. (1972). *Probability, Induction, and Statistics*. Wiley, New York.

DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

DIACONIS, P. and FREEDMAN, D. (1988). On the problem of types. Technical Report 153, Dept. Statistics, Univ. California, Berkeley.

DIACONIS, P. and FREEDMAN, D. (1990). On the uniform consistency of Bayes estimates for multinomial probabilities. *Ann. Statist.* **18** 1317–1327.

DIACONIS, P. and FREEDMAN, D. (1991). Nonparametric binary regression with unbalanced data. Technical Report 291, Dept. Statistics, Univ. California, Berkeley.

EUBANK, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

FREEDMAN, D. (1973). Another note on the Borel–Cantelli lemma and the strong law with the Poisson approximation as a by-product. *Ann. Probab.* **1** 910–925.

GEMAN, S. and HWANG, C. R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

GILLILAND, D. C., HANNAN, J. and HUANG, J. S. (1976). Asymptotic solutions to the two state component compound decision problem, Bayes versus diffuse priors on proportions. *Ann. Statist.* **4** 1101–1112.

GRENANDER, U. (1981). *Abstract Inference*. Wiley, New York.

GHOSH, J. K., SINHA, B. K. and JOSHI, S. N. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.) **1** 403–456. Academic, New York.

HALL, P. (1987). Cross-validation and the smoothing of orthogonal series density estimators. *J. Multivariate Anal.* **21** 188–206.

HART, J. D. (1988). An ARMA type probability density estimator. *Ann. Statist.* **16** 842–855.

HOEFFDING, W. (1956). On the distribution of the number of successes in independent trials. *Ann. Math. Statist.* **27** 713–721.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

JOHNSON, R. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Statist.* **38** 1899–1906.

JOHNSON, R. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Statist.* **41** 851–864.

LAPLACE, P. S. (1774). Memoire sur la probabilité des causes par les évènements. *Memoires de mathématique et de physique presentés a l'académie royale des sciences, par divers savants, et lûs dans ses assemblées* **6**. [Reprinted in *Laplace's Oeuvres Complètes* **8** 27–65. English translation by S. Stigler (1986) *Statist. Sci.* **1** 359–378.]

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.

LEONARD, T. (1978). Density estimation, stochastic processes, and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 113–146.

LI, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–104.

PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces*. Academic, New York.

RUDERMO, M. (1982). Empirical choice of histogram and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

SHIBATA, R. (1986). Consistency of model selection and parameter estimation. In *Essays in Time Series and Allied Processes* (J. Gani and M. B. Priestley, eds.). *J. Appl. Probab.* special vol. **23A** 127–141.

STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

DEPARTMENT OF MATHEMATICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720